Learning to be safe, in finite time

Agustin Castellano,[1] Juan Bazerque,[1] and Enrique Mallada[2]

Abstract— **This paper aims to put forward the concept that learning to take safe actions in unknown environments, even with probability one guarantees, can be achieved without the need for an unbounded number of exploratory trials, provided that one is willing to relax its optimality requirements mildly. We focus on the canonical multi-armed bandit problem and seek to study the exploration-preservation trade-off intrinsic within safe learning. More precisely, by defining a handicap metric that counts the number of unsafe actions, we provide an algorithm for discarding unsafe machines (or actions), with probability one, that achieves constant handicap. Our algorithm is rooted in the classical sequential probability ratio test, redefined here for continuing tasks. Under standard assumptions on sufficient exploration, our rule provably detects all unsafe machines in an (expected) finite number of rounds. The analysis also unveils a trade-off between the number of rounds needed to secure the environment and the probability of discarding safe machines. Our decision rule can wrap around any other algorithm to optimize a specific auxiliary goal since it provides a safe environment to search for (approximately) optimal policies. Simulations corroborate our theoretical findings and further illustrate the aforementioned trade-offs.**

I. Introduction

Learning to take *safe actions* in unknown environments is a general goal that spans across multiple disciplines. Within control theory safety is intrinsic to robust analysis and design [1], where controllers, with stability and performance guarantees, are designed for uncertain systems. It also at the core of statistical decision theory [2], where the inference and the decision-making processes are intertwined towards the common goal of making accurate decisions based on limited information. Though, historically, these two disciplines have been deemed as seemingly disconnected, such separation is rapidly vanishing.

Motivated by the success of machine learning in achieving super human performance, e.g., in vision [3], [4], speech [5], [6], and video games [7], [8], there has been recent interest in developing learning-enabled technology that can implement highly complex actions for safety-critical autonomous systems, such as self-driving cars, robots, etc. However, without proper safety guarantees such systems will rarely be deployed. There is therefore the need to develop analysis tools and algorithms that can provide such guarantees during, and after, training.

A common approach to solve this problem is to search for actions, policies, or controllers that optimize a cost or reward subject to safety requirements imposed as constraints. Examples include, adding safe constraints for reinforcement learning algorithms [9]–[13], (robust) stability constraints to learning algorithms [14]–[16], and solving constrained multi-armed bandits [17], [18]. When such constraints are being included during training, algorithms that converge to optimal policies, which are by definition feasible, guarantee safety asymptotically. However, such approach can only provide safety guarantees *after training*, and fails to provide guarantees in the exploratory period.

In this paper, we suggest an alternative approach. Instead of focusing on finding optimal actions subject to, a priori unknown, safety constrains, we argue that one should tackle the problem of learning safe actions separately and more efficiently. To illustrate this point, we study the problem of finding safe actions within the canonical setting of the multi-armed bandit problem, in which one is given a set of N machines/actions with expected reward $\mu_n \in [0,1]$ for $n \in \{1,\ldots,N\}$. By letting safe actions to be those who choose machines with reward larger than some nominal μ, we define a *handicap* metric —akin to regret— that counts the number of times unsafe actions are chosen.

Leveraging classical results on sequential hypothesis tests [19], we provide an algorithm for detecting unsafe machines. Unlike the regret minimization counterpart of this problem, which requires an unbounded (with logarithmic growth) number of trials of sub-optimal actions to identify the machine with highest reward [20], our algorithm discards all unsafe machines with only a finite number of trials of unsafe actions. By characterizing such number, we guarantee that the total handicap remains bounded by a constant.

More precisely, we use a modified version of Wald's sequential probability ratio test (SPRT) [19] for each machine, with the null hypothesis (H_0) being that the machine is safe and the alternative hypothesis (H_1) being that the machine is unsafe. Intentionally different than the SPRT, which aims to decide either H_0 or H_1 and then stop, our goal is to discard unsafe machines only by deciding on H_1. This allows our algorithm to *identify all unsafe machines with probability one* within a finite expected number of trials. Our analysis also unveils an *exploration-preservation trade-off* between the false-positive ratio (safe machines discarded) and the total handicap experienced (number of trials on unsafe machines). Notably, our decision rule can further wrap around any other algorithm to optimize a specific auxiliary goal since it provides a safe environment to search for (approximately) optimal policies.

The rest of the paper is organized as follows. We introduce our problem setup in Section II. For didactic purposes, we first look at the case where we aim to find machines with $\mu = 1$ in Section III, construct a sequential test that extends

[1] Agustin Castellano and Juan Bazerque are with the Universidad de la Republica, Montevideo, Uruguay. {acastellano, jbazerque}@fing.edu.uy. [2] Enrique Mallada is with the Johns Hopkins University, Baltimore, Maryland, USA. mallada@jhu.edu
The work was supported by NSF through grants CNS 1544771, EPCN 1711188, AMPS 1736448, and CAREER 1752362.

this case for one machine in Section IV, and generalize the solution in Section V. Numerical illustrations are provided in Section VI and we conclude in Section VII.

II. Problem statement

Consider the setup of a multi-armed bandit problem, in which at each time instant $t = 1, 2, \ldots$ we have the choice to operate one out of N machines. If machine $n \in \{1, 2, \ldots, N\}$ is operated at time t, it returns a binary value $X_{n,t}$ which is modelled as a Bernoulli random variable with parameter μ_n. This return reveals whether the action led to a safe result in which case $X_{n,t} = 1$, or an unsafe one if $X_{n,t} = 0$. A machine is said to be safe if its operation leads to a safe result. In this sense we consider two cases, one in which we only accept flawless machines, i.e., those with $\mu_n = 1$, and a relaxed condition in which a machine is defined to be safe if $\mu_n \geq \mu$, with $\mu \in (0, 1)$ being a prescribed safety requirement (possibly $\mu \simeq 1$).

Let $I_t \in \{1, 2, \ldots, N\}$ denote the index of the machine selected at time t, and $X_t \doteq X_{I_t,t}$ the corresponding return. Our goal is to design a selection policy and a decision rule that uses data $X_t,\ t = 1, 2, \ldots$ to remove all unsafe machines, while guaranteeing that a prescribed proportion of the safe machines are kept. If only flawless machines are accepted then the solution is straightforward: the algorithm should remove machines as soon as they return their first $X_t = 0$. We will analyse this case first in Section III.

For the relaxed condition, we will develop a one-sided Sequential Probability Ratio Test (SPRT) that removes unsafe machines with $\mu_n \leq \mu$ almost surely. In order to guarantee that unsafe machines are removed in finite time, and provide an explicit bound on the expected number of trials needed, we need to sacrifice a proportion of the safe machines. For this purpose we prescribe a slack parameter ϵ and a probability α, and show that a proportion $1 - \alpha$ of those safe machines with $\mu_n \geq \mu + \epsilon$ are kept as $t \to \infty$. We will develop this modified version of the SPRT in Section IV for the case of $N = 1$, extending it in Section V to the multi-armed bandit setup.

Along the way, we will introduce three figures of merits that are instrumental to goal of learning to be safe. One is the *handicap*, that complements the idea of regret for operating unsafe machines, and counts the number of unsafe actions chosen so far. Closely related to the notion of handicap is the testing time, that counts the number of times a machine is tried for safety, and is related to the detection time of unsafe machines. The third one is the safety ratio, which counts the proportion of safe machines that are kept at time t.

III. Safe learning with flawless machines

Consider the multi armed bandit setup described above with N machines, M of them unsafe or *malfunctioning*. In order to simplify notation and without loss of generality, we assume that the first M machines are unsafe so that $\mu_n < 1$ for $n = 1, \ldots, M$, and $\mu_n = 1$ for $n = M + 1, \ldots, N$.

We are assured that $X_{n,t} = 1\ \forall t$ if the machine is safe, thus we can discard those machines that return $X_{n,t} = 0$. This is the strategy in Algorithm 1, which selects actions at random over the set $\mathcal{S}_t$ of machines that remain at time t.

Algorithm 1: Safety Inspector

```
Initialize $\mathcal{S}_1 = \{1, \ldots, N\}$
for $t = 1, 2, \ldots,$ do
    Pick an arm $I_t \sim$ Unif($\mathcal{S}_t$)
    Observe return $X_t = X_{I_t,t}$
    if $X_t = 0$ then
        $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1} \setminus \{I_t\}$
    end
    else
        $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1}$
    end
end
```

The following definitions are introduced for the purpose of analysing the Safety Inspector Algorithm 1 and its relaxed version in Section V. First, even if we recognize that detecting unsafe machines requires unsafe actions to be taken, we want to measure if our algorithms pick those unsafe machines efficiently. For this purpose we present the notion of *handicap*, defined as the number of times an unsafe machine is selected, i.e.,

$$\text{Handicap}_t = t - \sum_{\tau=1}^{t} \mathbb{E}\left[\mathbb{1}\{\mu_{I_\tau} \geq \mu\}\right] \tag{1}$$

where $\mu = 1$ in this section, and $\mathbb{1}(\cdot)$ represents the indicator function which returns one or zero when its argument is true or false, respectively.

Remark. We use the word *handicap* in the sense of a measure of *"a disadvantage that makes achievement unusually difficult,"* as its definition suggests [21]. An algorithm with unbounded Handicap_t takes unsafe actions infinitely often and is prone to malfunctioning. This marks a stark contrast with the notion of *regret*, typically studied in Bandit settings [22], where unbounded regret is unavoidable [20].

Notice that if machine n is selected at time τ, then $\mu_{I_\tau} = \mu_n$. Thus, the indicator function will return 1 at time τ only when a flawless machine is selected. Furthermore, even if μ_n is deterministic, Handicap_t is still a random variable, with randomness coming from the selection I_t. Even if I_t is selected in round-robin instead of uniformly as in Algorithm 1, the set $\mathcal{S}_t$ is conditioned on previous instances of X_t, which are uncertain. In light of this, it is noticeable that an algorithm with low handicap is one which selects unsafe machines infrequently. As a second figure of merit we define the safety ratio ρ_t as the proportion of safe machines that are kept after t time slots, i.e.,

$$\rho_t = \frac{\sum_{n \in \mathcal{S}_t} \mathbb{1}(\mu_n \geq \mu + \epsilon)}{N - M} \tag{2}$$

with $\epsilon = 0$ for Algorithm 1. Together with the notions of handicap and safety ratio, we are interested in analysing the time that elapses until an unsafe machine is removed. For this purpose it is instrumental to define the number of times that machine n has been tested for safety after t iterations of Algorithm 1, that is

$$T_n(t) = \sum_{\tau=1}^{t} \mathbb{1}(I_\tau = n) \tag{3}$$

.

Next, we present a lemma that links the definitions of Handicap_t with $T_n(t)$, and will be useful to bound the expected handicap for Algorithm 1 and that in Section V.

Lemma 1. $\mathbb{E}[\text{Handicap}_t] = \sum_{n=1}^{M} \mathbb{E}[T_n(t)]$

Proof.

$$\begin{aligned}
&\mathbb{E}[\text{Handicap}_t] = \sum_{\tau=1}^{t} \mathbb{E}\left[\mathbb{1}\{\mu_{I_\tau} < \mu\}\right] \\
&= \sum_{\tau=1}^{t}\sum_{n=1}^{N} P(I_\tau = n)\mathbb{1}\{\mu_{I_\tau} < \mu\} = \sum_{\tau=1}^{t}\sum_{n=1}^{M} P(I_\tau = n) \\
&= \sum_{n=1}^{M}\sum_{\tau=1}^{t} \mathbb{E}\left[\mathbb{1}\{I_\tau = n\}\right] = \sum_{n=1}^{M} \mathbb{E}[T_n(t)]
\end{aligned}$$

□

Using the result in Lemma 1 we can bound the expected handicap of Algorithm 1 by bounding $\mathbb{E}[T_n(t)]$. This is the result of the next Theorem.

Theorem 1. *The handicap and safety ratio of Algorithm 1 satisfy*

$$\mathbb{E}[\text{Handicap}_t] \leq \sum_{n=1}^{M} \frac{1}{(1-\mu_n)^2} \tag{4}$$

$$\mathbb{E}[\rho_t] = 1, \tag{5}$$

and the testing time of unsafe machines is bounded by

$$\mathbb{E}[T_n(t)] \leq \frac{1}{(1-\mu_n)^2} \tag{6}$$

Proof. The result for the safety ratio is straightforward since the probability of removing a machine with $\mu_n = 1$ is zero. The bound for the handicap follows from Lemma 1, together with the result for the testing time, which is proved next

$$\mathbb{E}[T_n(t)] = \sum_{\tau=1}^{t} \mathrm{P}(T_n(t) = \tau)\tau \leq \sum_{\tau=1}^{t} \mu_n^{\tau-1}\,\tau \tag{7}$$

$$\leq \sum_{\tau=1}^{\infty} \mu_n^{\tau-1}\,\tau \leq \frac{1}{(1-\mu_n)^2} \tag{8}$$

□

Remark. The right hand side of (6) in Theorem 1 bounds the expected number of times that an unsafe machine is tested before removing it. This highlights one of the main ideas introduced in this paper: if we only want to detect unsafe machines instead of estimating the exact value of μ_n, then we can do it in finite time. As a consequence, the measure of handicap defined in (1) remains bounded by a constant. Even if we deem this result as conceptually relevant, it presents the drawback that the bounds for the expected handicap and testing times are given in terms of μ_n which are unknown. In order to provide an explicit bound in terms of the design parameters of the algorithm it is convenient to relax the condition that defines a safe machine, allowing for machines with $\mu_n \geq \mu$ lowering the prescribed safety threshold to $\mu < 1$. By doing so, we will retain the ability of rejecting all unsafe machines almost surely, while explicitly bounding the handicap. With this goal in mind, we present our modified SPRT in the next section.

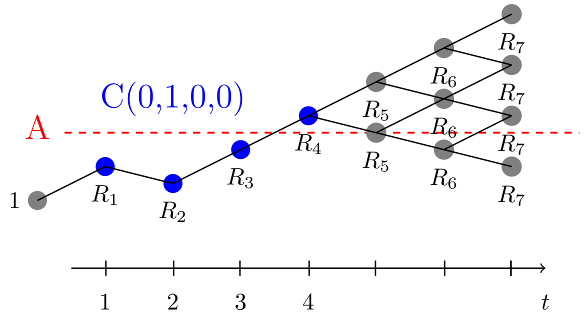

Fig. 1. Schematic of the decision rule for the one-sided SPRT (11). Sequences $\{x_t\}_{t=1}^{\infty}$ belonging to the cylinder set $C(0,1,0,0)$ coincide in the first $t = 4$ entries $(x_1, x_2, x_3, x_4) = (0, 1, 0, 1)$, which determine the likelihood ratios (L_1, L_2, L_3, L_4) represented by blue points. Since they cross the decision threshold A at $T = 4$, the null hypothesis is rejected and the machine is declared unsafe. The decision is made at $T = 4$, and therefore the multiple possible trajectories of L_t afterwards are inconsequential and can be disregarded.

IV. Sequential probability ratio test

Consider in this section the case of a single machine with unknown mean μ_n. We face the problem of deciding whether the machine is unsafe, i.e., $\mu_n \leq \mu < 1$. For this purpose, we set the following Hypothesis test

$$\begin{cases} H_0: & \mu_n \geq \mu + \epsilon \\ H_1: & \mu_n \leq \mu \end{cases} \tag{9}$$

where $\epsilon \leq 1 - \mu$ is a slack parameter. The goal of this section is to devise a sequential test, which uses data X_t for $t = 1, 2, \ldots$ to detect if the machine is unsafe. We look for a test that detects such a machine almost surely, and that guarantees that a machine with $\mu_n \geq \mu + \epsilon$ is kept with probability $1-\alpha$ as t grows unbounded. The three values μ, ϵ and α are design parameters. An overly conservative choice, $\epsilon \simeq 0$ $\alpha \simeq 0$ pays the price of a longer detection time, as shown in Lemma 4 later in this section. The construction of the following test and the analysis of cylinder sets are based on Wald's celebrated SPRT [19].

Let $f_\mu(x)$ and $f_{\mu+\epsilon}(x)$ denote the probability mass functions corresponding to the Bernoulli distributions of parameters μ and $\mu+\epsilon$, respectively. For $t \in \mathbb{N}$ consider a sampled trajectory $(x_1, \ldots, x_t)$ and define the likelihood ratio:

$$L_t = \frac{f_\mu\left(x_1, x_2, \ldots, x_t\right)}{f_{\mu+\epsilon}\left(x_1, x_2, \ldots, x_t\right)} \tag{10}$$

At each time we calculate L_t and accept H_1 if

$$L_t \geq A \tag{11}$$

where A is a design parameter that will be specified later. This is, if $L_t \geq A$ we declare that the machine is unsafe and stop the test. Otherwise we take an additional observation pulling the arm one more time to then check the condition (11) again after updating $t \to t+1$. Assuming i.i.d. samples, (11) can be transformed into a condition on the number of zeros k_t in the trajectory $(x_1, x_2, \dots, x_t)$. Indeed, taking the logarithm of (10), (11) transforms into $k_t\lambda_0 - (t-k_t)\lambda_1 \geq \log A$, with $\lambda_0 := \log(f_\mu(0)/f_{\mu+\epsilon}(0))$ and $\lambda_1 := -\log(f_\mu(1)/f_{\mu+\epsilon}(1))$. Rearranging terms we arrive to the equivalent condition for (11)

$$k_t \geq \frac{\log A + \lambda_1 t}{\lambda_0 + L1} \tag{12}$$

with

$$\lambda_0 = \log((1-\mu)/(1-\mu-\epsilon)) \tag{13}$$

$$\lambda_1 = \log((\mu+\epsilon)/\mu) \tag{14}$$

In this new form, it is apparent that our decision rule reduces to a sequential binomial test. Different from Algorithm 1, (12) does not discard a machine on the first zero, but has a probabilistic rule to decide when the number of zeros does not correspond with the hypothesis of a safe machine.

There are three questions that we would like to answer in this setting: i) do unsafe machines produce sequences that escape the threshold A with probability one?, ii) do safe machines produce sequences that do not escape this threshold, and if so, with what probability?, and iii) what is the expected time for the probability ratio of sequences coming from an unsafe machine to cross the threshold A? To answer these questions we need some definitions first.

A. Cylinder sets

Consider an infinite sequence $\{x_\tau\}, \tau = 1, 2, \dots$ and define C_∞ as the space of all such sequences. The set $C(a_1, \dots, a_t)$ is called a cylinder set of order t, and is defined as the subset of C_∞ which collects sequences with $x_1 = a_1, \dots, x_t = a_t$. A cylinder set will be said to be of the unsafe type if

$$L_t = \frac{f_\mu\left(a_1, \dots, a_t\right)}{f_{\mu+\epsilon}\left(a_1, \dots, a_t\right)} \geq A \tag{15}$$

and if for all $\tau = 1, \cdots, t-1$.

$$L_\tau = \frac{f_\mu\left(a_1, \dots, a_\tau\right)}{f_{\mu+\epsilon}\left(a_1 \dots, a_\tau\right)} < A \tag{16}$$

The first condition ensures that all infinite sequences $\{x_\tau\}$ with $x_1 = a_1, \dots, x_t = a_t$ will lead to the acceptance of hypothesis H_1, thus declaring the machine unsafe. This is depicted in Figure 1, showing a sequence that belongs to an unsafe cylinder of order $t = 4$. Notice that by construction, a machine that produces a sequence belonging to an unsafe cylinder as in (15) will be declared unsafe at time t, regardless of the future samples x_τ, $\tau > t$. Thus, we can effectively stop the test for that machine at time t. The second condition (16) ensures that cylinders of different orders are disjoint sets, since the probability ratio must exceed the threshold A for the first time at t an this cannot be true for two different values of t. The union of all unsafe cylinder sets (of any order) defines the set of sequences that lead to deciding H_1. Let us name this (disjoint) union as Q_U. Let us also define Q_S as the complement of Q_U

$$Q_S = Q_U^{\complement} \tag{17}$$

This definition means that Q_S is the set of all sequences for which the likelihood ratio L_t never rises above A. Because they are complementary, it holds for all $\mu_n \in [0,1]$

$$P_{\mu_n}(Q_U + Q_S) = 1 \tag{18}$$

with $P_{\mu_n}(Q)$ being the probability measure corresponding to a Bernoulli distribution of parameter μ_n.

We would like to obtain the following behavior:

- Under H_0, *most* sequences belong to Q_S
- Under H_1, *all* sequences belong to Q_U.

The second claim is guaranteed by the following lemma, which departs from [19] because there is a nonzero probability of not stopping the test, and thus requires a special treatment.

Lemma 2. *Let $P_{\mu_n}(Q)$ and $f_{\mu_n}(x)$ be the probability measure and mass function corresponding to a Bernoulli distribution of parameter μ_n under the alternative hypothesis H_1 ($\mu_n \leq \mu$) . Then, i.i.d. sequences produced by such a distribution are correctly classified almost surely, that is*

$$P_{\mu_n}(Q_U) = 1, \quad \forall\, \mu_n \leq \mu \tag{19}$$

Proof. For a given sequence $(x_1, \dots, x_t)$ define the log-likelihood ratio as

$$\Lambda_t = \log L_t = \log \prod_{i=1}^{t} \frac{f_\mu(x_i)}{f_{\mu+\epsilon}(x_i)} = \sum_{i=1}^{t} \log \frac{f_\mu(x_i)}{f_{\mu+\epsilon}(x_i)} \tag{20}$$

Dividing by t:

$$\frac{\Lambda_t}{t} = \frac{1}{t}\sum_{i=1}^{t} \log \frac{f_\mu(x_i)}{f_{\mu+\epsilon}(x_i)}$$

Taking the limit as $t \to \infty$ the above expression converges to the expectation of the right hand side under the alternative hypothesis

$$\begin{aligned}
\frac{\Lambda_t}{t} &\to \mathbb{E}_{x\sim f_{\mu_n}}\left[\log \frac{f_\mu(x)}{f_{\mu+\epsilon}(x)}\right] \\
&= \mu_n \log \frac{\mu}{\mu+\epsilon} + (1-\mu_n)\log \frac{1-\mu}{1-\mu-\epsilon} \\
&= \mu_n \left(\log \frac{\mu}{\mu+\epsilon} - \log \frac{1-\mu}{1-\mu-\epsilon}\right) + \log \frac{1-\mu}{1-\mu-\epsilon} \\
&\geq \mu \left(\log \frac{\mu}{\mu+\epsilon} - \log \frac{1-\mu}{1-\mu-\epsilon}\right) + \log \frac{1-\mu}{1-\mu-\epsilon} \\
&= \mu \log \frac{\mu}{\mu+\epsilon} + (1-\mu)\log \frac{1-\mu}{1-(\mu+\epsilon)}
\end{aligned}$$

$$= D_{KL}(f_\mu \,||\, f_{\mu+\epsilon}) > 0 \tag{21}$$

where D_{KL} stands for the Kullback-Leibler divergence. The inequality holds because $\mu_n \leq \mu$ by hypothesis, and it multiplies the expression in brackets, which is negative. From the inequality in (21) and the Law of Large Numbers, it follows

$$\lim_{t\to\infty} \Lambda_t = \infty, \quad a.s.$$

Therefore there must exist a positive integer t for which Λ_t exceeds $\log A$, so that the sequence $\{x_\tau\}$ belongs to an unsafe cylinder of order t and thus $\{x_\tau\} \in Q_U$. □

The previous Lemma proved that unsafe machines are detected with probability one. Next we prove that, by designing the threshold A judiciously, a fraction $1 - \alpha$ of the safe machines are kept in the system indefinitely.

Lemma 3. *Let $A = \frac{1}{\alpha}$ and $\mu_n \geq \mu + \epsilon$. Then, the probability that a trajectory never rises above A is $P_{\mu_n}(Q_S) \geq 1 - \alpha$.*

Proof. First, we prove the claim $P_{\mu+\epsilon}(Q_S) \geq 1 - \alpha$ for the limiting case $\mu_n = \mu + \epsilon$, and then we generalize it for $\mu_n \geq \mu + \epsilon$. For $\mu_n = \mu + \epsilon$, the core of the proof relies in showing that

$$L_t \geq A \Longrightarrow P_\mu(Q_U) \geq A P_{\mu+\epsilon}(Q_U) \tag{22}$$

Once we prove (22), we use the fact that $P_\mu(Q_U) = 1$ (see (19)), hence

$$P_{\mu+\epsilon}(Q_U) \leq \frac{1}{A} = \alpha \Rightarrow P_{\mu+\epsilon}(Q_S) \geq 1 - \alpha \tag{23}$$

as desired.

To prove (22), we start by decomposing Q_U as the union across time t of the union of all unsafe cylinders of order t, that is

$$Q_U = \bigcup_{t=1}^{\infty} \bigcup_{(a_1,\ldots,a_t)\in\mathcal{A}_t} C(a_1,\ldots,a_t)$$

where $\mathcal{A}_t$ collects the tuples $(a_1,\ldots,a_t)$ that define unsafe cylinders of order t, i.e., those satisfying (15) and (16). By construction all the cylinder sets are disjoint, hence

$$\begin{aligned}
P_\mu(Q_U) &= \sum_{t=1}^{\infty} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}_t} P_\mu\left(C(a_1,\ldots,a_t)\right) \\
&= \sum_{t=1}^{\infty} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}_t} f_\mu(a_1,\ldots,a_t) \\
&\geq \sum_{t=1}^{\infty} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}_t} A f_{\mu+\epsilon}(a_1,\ldots,a_t) \\
&= A \sum_{t=1}^{\infty} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}_t} P_{\mu+\epsilon}(C(a_1,\ldots,a_t)) \\
&= A P_{\mu+\epsilon}(Q_U)
\end{aligned}$$

where second identity follows from marginalizing over future trajectories (see Fig. 1), and the inequality holds since $\mathcal{A}_t$ is defined to satisfy (15).

Now that we have (22), (23) follows immediately, and we need to prove that $P_{\mu_n}(Q_S) \geq P_{\mu+\epsilon}(Q_S)$, or equivalently $P_{\mu_n}(Q_U) \leq P_{\mu+\epsilon}(Q_U)$, when $\mu_n \geq \mu + \epsilon$. This is intuitively true, since for a sequence to belong to Q_U, it must satisfy (12). But the number of zeros is distributed $k_t \sim \text{Binomial}(t, 1-\mu_n)$, so that the probability of satisfying (12) becomes lower as μ_n increases. For a more rigorous proof, we decompose again

$$P_{\mu_n}(Q_U) = \sum_{t=1}^{\infty} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}_t} f_{\mu_n}(a_1,\ldots,a_t) \tag{24}$$

as we did for $P_\mu(Q_U)$. We will prove that for any fixed tuple $(a_1,\ldots,a_t) \in \mathcal{A}_t$, $f_{\mu_n}(a_1,\ldots,a_t)$ is a decreasing function of μ_n. First, we need to prove that the number of zeros k_t in $(a_1,\ldots,a_t)$ satisfies $1 - k_t/t \leq \mu_n$. But because k_t in $(a_1,\ldots,a_t)$ must satisfy (12), and $\log A$, λ_0, and λ_1 are strictly positive, then $k_t \geq \lambda_1 t/(\lambda_0 + \lambda_1)$ or equivalently $1 - k_t/t \leq \lambda_0/(\lambda_0 + \lambda_1)$. Form the definition of λ_0 and λ_1 it yields

$$\begin{aligned}
1 - \frac{k_t}{t} \leq \frac{\lambda_0}{\lambda_0 + \lambda_1} &= \frac{\log[(1-\mu)/(1-\mu-\epsilon)]}{\log[(\mu+\epsilon)(1-\mu)/(\mu(1-\mu-\epsilon))]} \\
&\leq \frac{(1-\mu)/(1-\mu-\epsilon) - 1}{1 - \mu(1-\mu-\epsilon)/((\mu+\epsilon)(1-\mu))} \qquad (25) \\
&= (\mu+\epsilon)(1-\mu) \leq \mu + \epsilon \leq \mu_n \qquad (26)
\end{aligned}$$

where the first inequality results from the usual bounds of the logarithm $1 - 1/x \leq \log(x) \leq x - 1$. Rearranging $1 - \frac{k_t}{t} \leq \mu_n$, it results $t - k_t - t\mu_n \leq 0$. In this case, the derivative of $f_{\mu_n}(a_1,\ldots,a_t)$ takes the form

$$\begin{aligned}
\frac{d}{d\mu_n} f_{\mu_n}(a_1,\ldots,a_t) &= \frac{d}{d\mu_n} \mu_n^{t-k_t}(1-\mu_n)^{k_t} \\
&= \mu_n^{t-k_t-1}(1-\mu_n)^{k_t-1}(t - k_t - t\mu_n) \leq 0 \qquad (27)
\end{aligned}$$

Putting (18), (23), (24), and (27) together results in $P_{\mu_n}(Q_S) \geq 1 - \alpha$ for all $\mu_n \geq \mu + \epsilon$. □

We have answered two of the three questions about our one-sided SPRT. Once we know that all unsafe machines are detected with probability one, it remains to characterize the detection time, which is the purpose of the following lemma. Henceforth, we will set the decision threshold at $A = \frac{1}{\alpha}$.

Lemma 4. *Under the alternative hypothesis H_1 corresponding to $\mu_n \leq \mu$, and with $A = \frac{1}{\alpha}$, the test* (11) *is expected to terminate after T steps, with*

$$\mathbb{E}[T] \leq 1 + \frac{\log\left(1/\alpha\right)}{D_{KL}(f_\mu \,||\, f_{\mu+\epsilon})} \tag{28}$$

Proof. Let T be the smallest integer for which the test leads to the acceptance of H_1. Such variable is well defined and finite as a result of Lemma 2.

$$\mathbb{E}_{x\sim f_{\mu_n}}\left[\Lambda_T\right] = \mathbb{E}_T\left[\mathbb{E}_{x\sim f_{\mu_n}}\left[\sum_{i=1}^{T} \log\frac{f_\mu(x_i)}{f_{\mu+\epsilon}(x_i)} \;\middle|\; T\right]\right]$$

$$\begin{aligned} &= \mathbb{E}_T \left[T \mathbb{E}_{x \sim f_{\mu_n}} \left[\log \frac{f_\mu(x)}{f_{\mu+\epsilon}(x)} \right] \right] \\ &= \mathbb{E}[T] \mathbb{E}_{x \sim f_{\mu_n}} \left[\log \frac{f_\mu(x)}{f_{\mu+\epsilon}(x)} \right] = \mathbb{E}[T] L_n \end{aligned} \tag{29}$$

with $L_n = \mathbb{E}_{x \sim f_{\mu_n}} \left[\log \frac{f_\mu(x)}{f_{\mu+\epsilon}(x)} \right]$. Furthermore,

$$\begin{aligned} \mathbb{E}_{x \sim f_{\mu_n}} [\Lambda_T] &= \mathbb{E}_{x \sim f_{\mu_n}} \left[\Lambda_{T-1} + \log \frac{f_\mu(x_T)}{f_{\mu+\epsilon}(x_T)} \right] \\ &= \mathbb{E}_{x \sim f_{\mu_n}} [\Lambda_{T-1}] + L_n \\ &\leq \log(1/\alpha) + L_n \end{aligned} \tag{30}$$

Combining (29) and (30):

$$\mathbb{E}[T] \leq 1 + \frac{\log\left(1/\alpha\right)}{L_n} \leq 1 + \frac{\log\left(1/\alpha\right)}{D_{KL}(f_\mu \mid\mid f_{\mu+\epsilon})}$$

in virtue of $L_n \leq D_{KL}(f_\mu \mid\mid f_{\mu+\epsilon})$ for $\mu_n \leq \mu$, as it was proved in (21). □

The result in Lemma 4 evidences the need of some slack ϵ between the limiting distributions of both hypotheses. By accommodating this gap we are able to separate the limiting distributions $f_\mu(x)$ from $f_{\mu+\epsilon}(x)$ so that the distance $D_{KL}(f_\mu \mid\mid f_{\mu+\epsilon})$ is positive and we can guarantee a finite expected detection time. Notice that we could use L_n for the bound on the detection time, as given in the proof of Lemma 4, which indeed gives a tighter bound, meaning faster detection. However, it requires the knowledge of the underlying probability μ_n which is unknown. Using the D_{KL} instead is preferred, because it yields a bound that depends on our design parameters μ, ϵ and α only. This result also allows us to have some intuitive interpretation regarding the choice of these design parameters.

With lemmas 2–4 at hand, we return to our original setup.

V. Learning to be safe

Next we generalize Algorithm 1 for the case in which the safety requirement $\mu = 1$ is relaxed. The following algorithm results from extending the one-sided SPRT just described to the scenario with multiple machines. Identical to the previous Section, we prescribe a safety threshold $\mu < 1$ that renders machines with $\mu_n < \mu$ as unsafe. Then, we define an error probability α, a slack parameter ϵ, and the Bernoulli probability mass functions $f_{\mu+\epsilon}(x)$ and $f_\mu(x)$ with means $\mu + \epsilon$ and μ respectively. These are all the definitions needed to run our second Safety Inspector algorithm.

Algorithm 2: Relaxed Safety Inspector

```
Initialize $\mathcal{S}_1 = \{1, \ldots, N\}$, $\Lambda_n = 0,\ n = 1, \ldots, N$
for $t = 1, 2 \ldots$ do
    Pick an arm $I_t \sim$ Unif($\mathcal{S}_t$)
    Observe return $X_t = X_{I_t, t}$
    Update $\Lambda_{I_t} += \log \frac{f_\mu(X_t)}{f_{\mu+\epsilon}(X_t)}$
    if $\Lambda_{I_t} \geq \log(1/\alpha)$ then
        $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1} \setminus \{I_t\}$
    end
end
```

Building on lemmas 1–4 we state our main result.

Theorem 2. *The handicap and safety ratio of the Relaxed Safety Inspector (Algorithm 2) satisfy*

$$\mathbb{E}[\text{Handicap}_t] \leq M \left(1 + \frac{\log\left(1/\alpha\right)}{D_{KL}(f_\mu \mid\mid f_{\mu+\epsilon})} \right) \tag{31}$$

$$\mathbb{E}[\rho_t] \leq 1 - \alpha \tag{32}$$

and the testing time of unsafe machines is bounded by

$$\mathbb{E}[T_n(t)] \leq 1 + \frac{\log\left(1/\alpha\right)}{D_{KL}(f_\mu \mid\mid f_{\mu+\epsilon})} \tag{33}$$

Proof. The third inequality was proved in Lemma 4. Notice that $T_n(t)$ is defined as the number of trials for machine n, regardless of the time spent on other machines, so that we can treat it as the result of separate SPRTs, and thus use Lemma 4. The first inequality follows from (33) and Lemma 1. The second one results from Lemma 3. □

Theorem 2 certifies that the Relaxed Safety Inspector (Algorithm 2) inherits the finite detection time from the SPRT, ensuring that all unsafe machines are removed in finite time, providing a universal bound (33) in terms of the design parameters α, μ and ϵ. As a consequence, the total handicap of the system also remains bounded by a finite constant (31). Together with the certainty of rejecting all unsafe machines, our algorithm ensures that a proportion $1 - \alpha$ of the safe machines with slack ϵ is kept in the system indefinitely according to (32).

VI. Numerical examples

Let us first illustrate the behavior of Algorithm 2 and how it proceeds to discard unsafe arms. To that end, consider a simple setup of $N = 3$ arms, all with same parameter $\mu_n = 0.8$. Our safety requirements are set to $\mu = 0.9$, $\epsilon = 0.02$ and the error probability to $\alpha = 0.05$. With this parameters in mind, all machines should be deemed unsafe in finite time. Our discarding rule involves checking when $\Lambda_t \geq \log A$. Since we are dealing with Bernoulli random variables, this rule can be equivalently cast as a decision based on the number of failed outcomes of each arm (12). These reciprocal ideas are depicted on the test in Figure 2.

A. Experiment 1: Transient behavior

We consider a setup of $N = 1000$ arms, and set a safety guarantee $\mu = 0.9$ and a gap $\epsilon = 0.05$. The true parameter of each arm is sampled from a uniform distribution $\mathcal{U}\left(0.8, 1\right)$. We run sixteen instances of the Safety Inspector described in Algorithm 2 on this test-bed, and then average the results obtained. We define the *normalized*

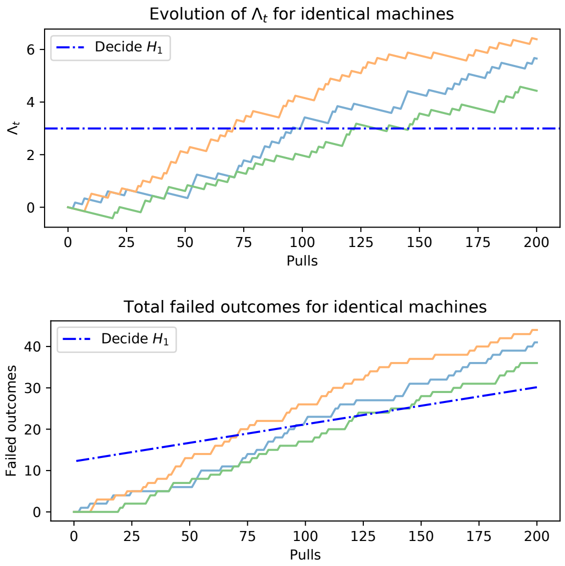

Fig. 2. Sequential probability ratio test for $\mu = 0.9$, $\epsilon = 0.02$, $\alpha = 0.05$ and three identical unsafe machines of parameter $\mu_n = 0.8$. Above: log-likelihood function Λ_t as a function of the number of pulls of each machine. Below: failed outcomes for each machine as a function of the number of pulls. The test terminates when the sequence surpasses either rejection line (in blue). All three unsafe machines are discarded in finite time.

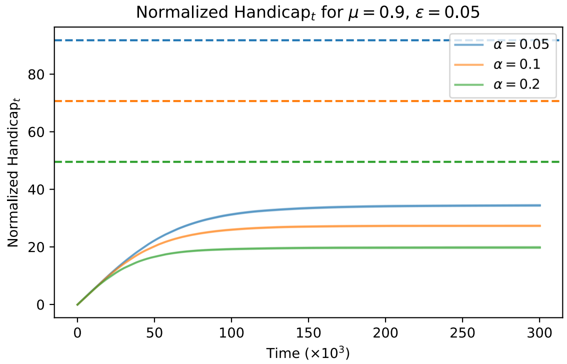

Fig. 3. Evolution of the Normalized Handicap through training for $\mu = 0.9$, $\epsilon = 0.05$ and machines with parameter drawn from $\mathcal{U}(0.8, 1)$. Each solid line corresponds to the handicap obtained with a different error tolerance α, and the dashed lines are the (normalized) bounds on the Handicap (see (31)). All unsafe machines are eventually discarded after sufficient training, and therefore the handicap remains constant.

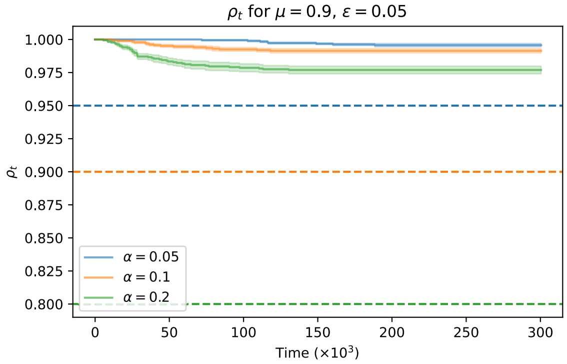

Fig. 4. Evolution of the safety ratio ρ_t through training for $\mu = 0.9$, $\epsilon = 0.05$ and machines with parameter drawn from $\mathcal{U}(0.8, 1)$, for different tolerance levels α. Each solid line is accompanied by its corresponding bound (see (32)).

handicap as the average handicap over all the available arms: $\text{NHandicap}_t = \frac{1}{N}\text{Handicap}_t$. Figures 3 and 4 show the evolution of the normalized handicap and safety ratio for different tolerance levels α, along with the bounds obtained in Theorem 2. Notice that both the handicap and safety ratio remain constant after some time, which indicates that all unsafe machines have been identified, and that no more safe machines are discarded along the way. The final handicap obtained is essentially the number of pulls over all unsafe arms. This is depicted more closely in Figure 5, which presents a histogram of the *testing time* on unsafe machines, for the setup explained above and for fixed $\alpha = 0.1$. Most machines yield a testing time that is strictly lower than the bound in (8). It is important to remark that this bound is on the *expected* time, and therefore a small number of machines actually need to be tested for longer. Nevertheless, the empirical mean represented by a dashed green line in Fig. 5 lies below the red line that represents the bound. This bound, as well as those in Figs. 3 and 4, is loose because the machines parameters μ_n were drawn uniformly from $\mathcal{U}\left(0.8, 1\right)$, but becomes tight if selected as the limiting parameters of the hypothesis test $\mu_n \in \{\mu, \mu + \epsilon\}$.

B. Experiment 2: Steady state behavior

Repeating the same setup as in *Experiment 1*, now we perform multiple runs for varying ϵ. Let us define NHandicap_∞ as the maximal normalized handicap obtained —that is, the normalized handicap when all unsafe machines are discarded. Figure 6 illustrates the dependence of NHandicap_∞ for varying ϵ and α. Larger values of α and ϵ attain lower handicap, which essentially means that unsafe arms are detected faster. This, however, comes at a price —faster detection necessarily implies discarding safe machines along the way. Figure 7 shows the other side of the coin: the final value of the safety ratio ρ_t after all unsafe machines have been discarded, which we dub ρ_∞. As ϵ and α grow, ρ_∞ diminishes. The conjunction of these two Figures exemplify the preservation-exploration trade-off inherent to our Algorithm.

VII. Conclusions

In this paper we are interested in providing a safety environment for learning. To that end, we advance the idea that detecting if an action is safe is much simpler than trying to estimate its value function, and can be done in finite time. In this direction, we define a measure of handicap that complements the notion of regret by accounting for the aggregate number of unsafe actions explored. We focus on the multi-armed bandit problem, with the goal of detecting the malfunctioning machines and keeping the handicap bounded. For this purpose, we introduced the Relaxed Safety Inspector (Algorithm 2), which we developed

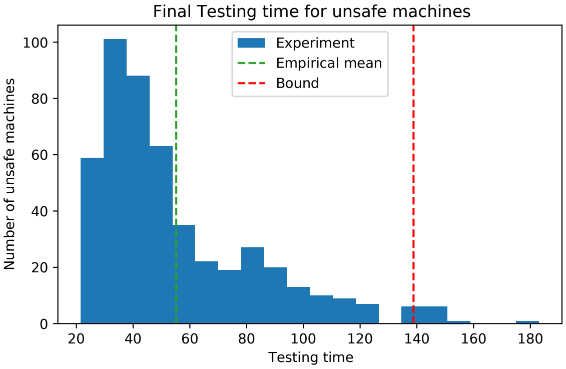

Fig. 5. Histogram of the testing time needed to discard unsafe machines, for $\mu = 0.9$, $\epsilon = 0.05$, $\alpha = 0.1$ and machines' parameters drawn from $\mathcal{U}(0.8, 1)$. In dashed red: bound on the testing time from Theorem 2. Since this bound is on the expected testing time, some machines need to be tested for longer. In dashed green: Empirical mean testing time, which is strictly lower than the bound described.

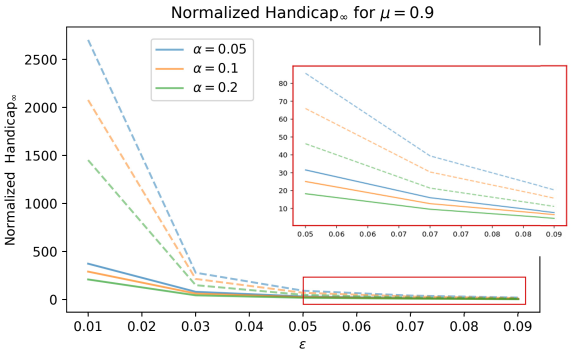

Fig. 6. Normalized final Handicap for $\mu = 0.9$ as a function of ϵ, for varying tolerance level α. As is to be expected, larger values of α and ϵ achieve lower handicap (which implies faster detection).

as sequential probability ratio test for parallel hypotheses. We proved in Theorem 2 that this algorithm has the property of removing all unsafe machines in finite time, providing a universal bound (33) in terms of the the design parameters α, μ and ϵ. As a consequence of this, the handicap remains bounded by a finite constant as time goes to infinity. The price to pay for being able to detect all malfunctioning machines in finite time is to accommodate a slack on the machines that are considered safe, and losing a proportion of them. Interestingly, these are design parameters that can be tightened if we are willing to wait longer for detection.

References

[1] K. Zhou, J. C. Doyle, K. Glover, *et al.*, *Robust and optimal control*, vol. 40. Prentice hall New Jersey, 1996.

[2] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[4] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

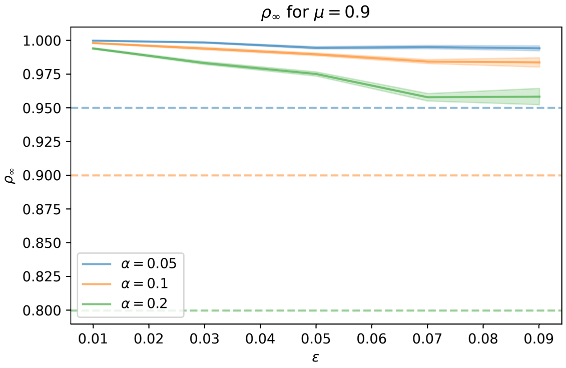

Fig. 7. Final safety ratio ρ_∞ for $\mu = 0.9$ as a function of ϵ, for varying α. More machines are kept when using small values of ϵ and α, but this in turn implies longer training time (c.f. Figure 6)

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, IEEE, 2013.

[7] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[8] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, *et al.*, "Starcraft ii: A new challenge for reinforcement learning," *arXiv preprint arXiv:1708.04782*, 2017.

[9] P. Geibel, "Reinforcement Learning for MDPs with Constraints," vol. 4212 of *Machine Learning: ECML 2006*, pp. 646–653, 2006.

[10] M. Zanon and S. Gros, "Safe Reinforcement Learning Using Robust MPC," *arXiv*, 2019.

[11] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, "Safe Policies for Reinforcement Learning via Primal-Dual Methods," *arxiv*, 2019.

[12] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks," *arXiv*, 2019.

[13] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained Policy Optimization," *arXiv*, vol. cs.LG, 2017.

[14] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the Sample Complexity of the Linear Quadratic Regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[15] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely Learning to Control the Constrained Linear Quadratic Regulator," *arxiv*.

[16] M. Fazlyab, M. Morari, and G. J. Pappas, "Safety Verification and Robustness Analysis of Neural Networks via Quadratic Constraints and Semidefinite Programming," *arXiv*, 2019.

[17] A. Moradipari, M. Alizadeh, and C. Thrampoulidis, "Linear Thompson Sampling Under Unknown Linear Constraints," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 00, pp. 3392–3396, 2020.

[18] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear Stochastic Bandits Under Safety Constraints," *arXiv*, 2019.

[19] A. Wald, "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, vol. 2, no. 16, pp. 117—186, 1945.

[20] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[21] Merriam-Webster, "Handicap."

[22] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.